BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 10: e183404
doi: 10.3897/biss.10.183404

OPEN ACCESS

Conference Abstract

Learning from Digital Collections: Digitisation and Machine Learning for Large-Scale Museum Studies

David R. Williamson‡, Emily A. Hartop‡, James D. M. Speed‡

‡ NTNU University Museum, Trondheim, Norway

Corresponding author: David R. Williamson (david.williamson@ntnu.no)

Received: 23 Dec 2025 | Published: 07 Jan 2026

Citation: Williamson DR, Hartop EA, Speed JDM (2026) Learning from Digital Collections: Digitisation and Machine Learning for Large-Scale Museum Studies. Biodiversity Information Science and Standards 10: e183404. https://doi.org/10.3897/biss.10.183404

Abstract

The world's museums collectively preserve billions of natural history specimens: identified by experts, curated, and verified. These specimens provide unique and irreplaceable historical data that can help to answer our most pressing questions relating to climate change, changes in phenology, habitat and species range, and biodiversity loss. Large-scale studies incorporating museum collections are vital to understanding these topics, but unlocking the information in collections typically requires physical access to the specimens and careful examination by experts. For the large-scale studies needed to understand global changes, involving millions of specimens across many nations, this is impractical. It is therefore increasingly important for museums to provide digital representations of specimens, comprising images and associated metadata, which can be accessed by researchers anywhere on the planet through searchable online databases. Digital specimens and modern computational analysis allow studies to be conducted at geographic and temporal scales that were previously impossible. We present two case studies demonstrating the value and limitations of such an approach.

Firstly, we investigated global flowering phenology using millions of digital specimens from museums and herbaria, applying machine learning (ML) to rapidly estimate phenology from images and reveal long-term, worldwide trends (Williamson et al. 2025). This global study aggregated data from museum collections spanning three centuries, leveraging collection and digitisation efforts by hundreds of herbaria, and using modern

© Williamson D et al. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

computational methods to examine phenological trends at an unprecedented scale. We identified local changes in flowering seasonality, opening the door to fine-grained analysis of the effects of climate change on the world's flora (Fig. 1). However, this study represents only ~2% of the estimated 400 million specimens held in the world's herbaria (Davis 2023). Lack of funding for digitisation, inadequate standardisation of digitisation methods, and inconsistent metadata and data standards remain a challenge for automated studies of this type, with more than 40% of records matching our search criteria failing data quality checks (see Williamson et al. (2025) for details).

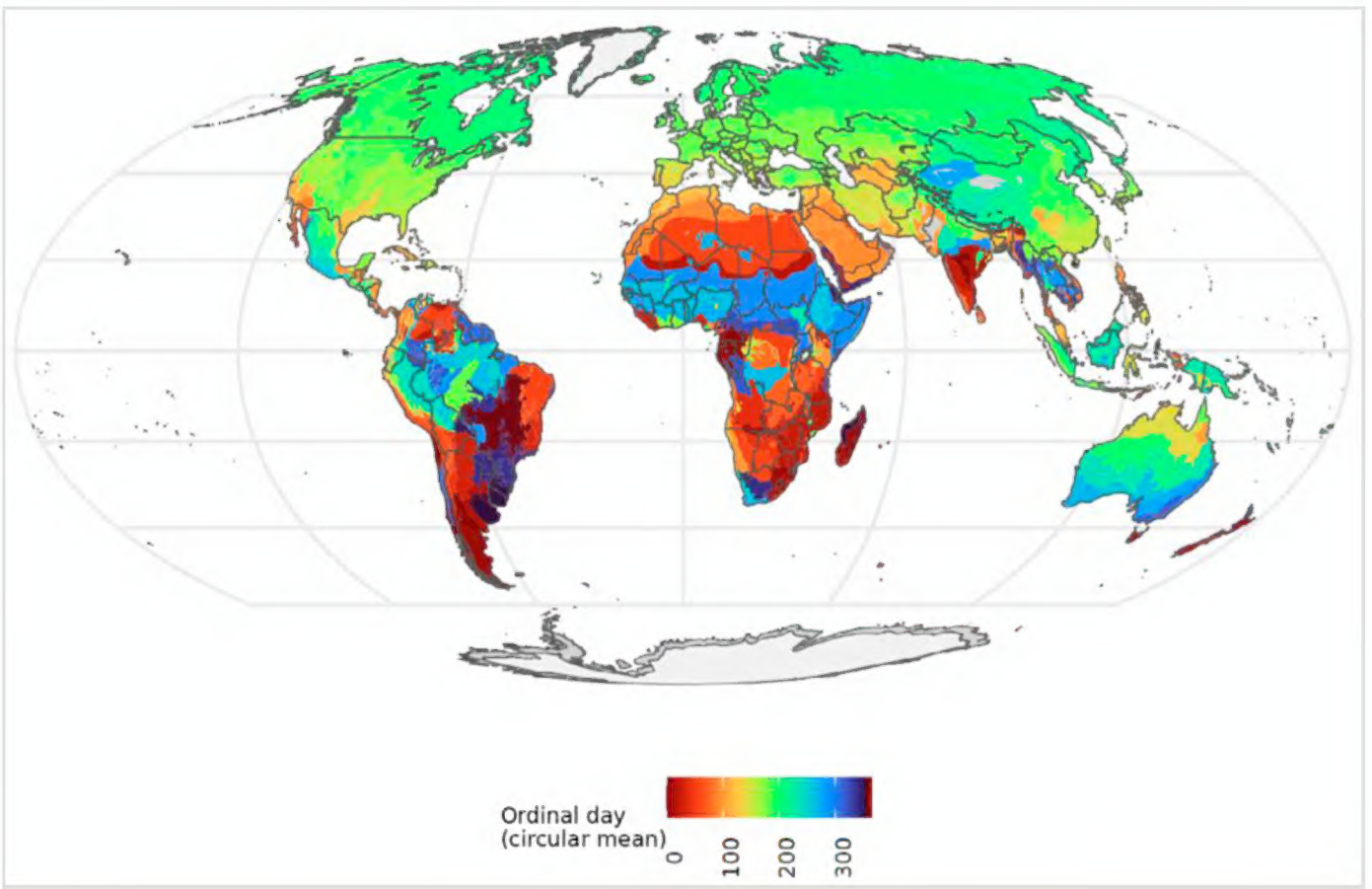

Figure 1.

Mean date of vascular plant flowering based on 8 million digitised herbarium sheets. Adapted with permission from Williamson et al. 2025. The author has obtained a personal, non-exclusive, non-sub licensable (on a stand-alone basis), non-transferable, worldwide, limited license to reproduce the Wiley Materials for the purpose specified in the licensing process.

In our second case study we digitised a portion of the R.H.L. Disney collection of scuttle flies (Diptera: Phoridae) at the Cambridge University Museum of Zoology. This fly family represents a "dark taxon" that may contain more than 100,000 species, of which <10% are described (Hartop et al. 2022). Dark taxa are both critical to measures of biodiversity and, by their hyperdiverse and cryptic nature, difficult to assess using traditional methods. The collection contains thousands of specimens, many of them type material, in some cases the only documented exemplar of a species. These specimens were mounted on glass slides making molecular methods such as DNA barcoding impossible. We used ML to extract morphometric landmarks to describe specimen characters and cluster specimens into species, allowing rapid identification of new specimens and letting the existing collection be tied to molecular analyses (Fig. 2). Since the collection was gathered over decades, with some specimens more than 50 years old, digitisation and

ML were not a consideration in their mounting or documentation, leading to difficulties including unstandardised metadata and degradation of specimens between collection and digitisation. While it is not possible to address such problems in historical collections, this highlights the need to ensure that future collection and digitisation efforts produce data that are directly machine actionable. We welcome work such as East et al. (2025), which aims to provide guidelines for standardised, effective digitisation.

Figure 2.

Phorid wing digitisation and first analysis workflow. Clockwise from top-left: collection in boxes, tray of specimen slides, individual slide of a dissected specimen, digitised wing with ML-predicted landmarks, a collection of Procrustes-normalised landmarks with colours indicating species. Figure by DR Williamson, licensed under CC BY 4.0.

By presenting these studies, and the challenges they involved, we aim to stimulate conversation on best practices around digitisation, data collection and dissemination, and the use of software tools for large-scale analysis of museum specimens to address pressing ecological questions.

Keywords

museum collections, phenology, species traits

Presenting author

David R. Williamson

Presented at

Living Data 2025

Acknowledgements

We thank Dr. Henry Disney and the Cambridge University Museum of Zoology for granting access to their phorid collection, and Dr. Mark Welch for his digitisation efforts.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Davis C (2023) The herbarium of the future. Trends in Ecology & Evolution 38 (5): 412-423. https://doi.org/10.1016/j.tree.2022.11.015
- East A, Campolongo EG, Meyers L, et al. (2025) Optimizing image capture for computer vision-powered taxonomic identification and trait recognition of biodiversity specimens. Methods in Ecology and Evolution 16 (10): 2260-2275. https://doi.org/10.1111/2041-210x.70140
- Hartop EA, Srivathsan A, Ronquist F, Meier R (2022) Towards Large-Scale Integrative Taxonomy (LIT): Resolving the Data Conundrum for Dark Taxa. Systematic Biology 71 (6): 1404-1422. https://doi.org/10.1093/sysbio/syac033
- Williamson DR, Prestø T, Westergaard KB, et al. (2025) Long-term trends in global flowering phenology. New Phytologist https://doi.org/10.1111/nph.70139